Разработчики ИИ всех стран, объединяйтесь!

еженедельная газета о новостях искусственного интеллекта

ИИ-ИЗВЕСТИЯ

№ 2026.11.23
еженедельный выпуск
Цена 2 коп.

Понедельник, 23 ноября 2026 года

4 полосы

Россия и мир · 16–23 ноября 2026

Главное в этом выпуске | **Карта выпуска** | **Новости и их значение** | **Что отслеживать**

Главное в этом выпуске

Неделя 16–23 ноября: ИИ берёт на себя рабочие поручения, модели получают новые возможности, а компании усиливают инфраструктуру и контроль агентов.

Новости подготовил Гермес — агент-обозреватель на навыках news-digest и news-monitor с проверкой источников через goap-research-ed25519; прочитал и отредактировал Дмитрий Жечков.

Карта выпуска

№	Событие	Дата
1	Яндекс подключает Алису и внешних агентов к сервисам.	16–20 ноября
2	MWS AI открыла CoSpaDi для сжатия языковых моделей.	20 ноября
3	Российский бытовой ИИ: локальный голос и разбор почты.	17 ноября
4	OpenAI выпустила GPT-6.2 Astra и Sol.	17 ноября
5	Google: Gemini 4 Neon для работы, Genie 4 для интерактивных миров.	18–19 ноября
6	Gemini начинает действовать на Mac и расширяет бытовые поручения.	16–20 ноября
7	Anthropic закрыла IPO.	20 ноября
8	Cohere добавляет поиск по видео и бизнес-данным.	16–17 ноября
9	Anthropic сокращает сроки исправления уязвимостей; независимая оценка становится ближе к разработке.	16–20 ноября
10	Microsoft Ignite: агенты входят в корпоративный рабочий контур.	17–20 ноября
11	NVIDIA: научные агенты, процессоры Vera и локальные RTX Spark.	16–18 ноября
12	Dell расширяет AI Factory, AMD покупает Taalas.	16–18 ноября
13	AWS связывает документы, речь и совместную работу агентов.	18–21 ноября
14	Cloudflare открыла Clairic, OVH — несколько вариантов Olala.	18–19 ноября
15	Инструменты агентов экономят память, токены и работу инженеров.	16–20 ноября

Что произошло

Период: 16 ноября 2026, 06:00 — 23 ноября 2026, 06:00 МСК

Главное.

Неделя сблизила ИИ с обычной работой: Яндекс 360 открыл подключение агентов к рабочим сервисам, а Gemini для Mac начал выполнять действия в приложениях. Между советом в чате и выполненным поручением становится меньше ручных переключений.

В моделях — два заметных направления. GPT-6.2 Astra и Sol разделяют тяжёлые профессиональные задачи и повседневную работу; Gemini 4 Neon соединяет большой контекст с агентскими сценариями. Genie 4 предлагает другой интерфейс: генерируемый мир, которым можно управлять, а не только смотреть готовое видео.

У развития ИИ появилась и финансовая веха: Anthropic закрыла IPO. Одновременно Microsoft Ignite и обновления AWS показали, что конкуренция идёт не только за интеллект модели, но и за доступ к данным, выполнение задач и управление агентами.

Россия: сервисы и локальные модели.

16–20 ноября —

Яндекс подключает Алису и внешних агентов к сервисам.

Яндекс 360 открыл предварительный доступ к MCP-серверу — стандартному соединению, через которое ИИ-агент получает инструменты рабочих сервисов. Поддерживаются Почта, Диск, Календарь, Телемост, Вики, Трекер и Формы. Агент действует с правами пользователя; в Почте может подготовить черновик, но не отправить письмо. Ранее на неделе Алиса AI получила бета-каталог более чем из 20 плагинов , включая возможность подключать внешние MCP-серверы и частные интеграции компаний.

Почему важно. Обычный чат-бот не знает содержимого рабочих папок и не может сам создать задачу или событие в календаре. Подключения Яндекса дают агенту доступ к конкретным данным и действиям без отдельной интеграции под каждую модель. Для сотрудников появляется возможность поручать сквозную задачу — найти материал, подготовить документ и завести задачу — вместо переноса информации вручную между приложениями.

20 ноября —

MWS AI открыла CoSpaDi для сжатия языковых моделей.

MWS AI опубликовала код CoSpaDi: метод уменьшает размер уже обученной языковой модели без дополнительного обучения.

Редакция сердечно благодарит за идею формата и пример реализации Станислава Гридина https://t.me/gridstan

Разработчики проверили подход на моделях Llama и Qwen и сообщили о преимуществе перед SVD-LLM по средней точности при сжатии на 20–40%. Результаты относятся к испытанным конфигурациям, а не ко всем моделям и задачам.

Почему важно. Крупная модель может не помещаться в доступную видеопамять, а её повторное обучение требует вычислений и данных. CoSpaDi сокращает модель после обучения, поэтому потенциально позволяет использовать имеющиеся серверы вместо покупки более дорогих ускорителей. Для компаний с собственным развёртыванием практический выигрыш зависит от того, насколько хорошо после сжатия сохраняются нужные им способности модели.

---

17 ноября —

Российский бытовой ИИ: локальный голос и разбор почты.

SberDevices начала продажи СберБум Трек. Локальная модель ГигаЧат распознаёт команды для управления сохранённой музыкой без интернета; автономность не означает работу всего диалогового ассистента без сети. В тот же день Почта Mail представила ИИ-виджеты и «Важные письма» — дайджест значимой переписки, первоначально в Android-супераппе.

Почему важно. Голосовые команды нередко перестают работать при плохой связи, а важные сообщения теряются среди рассылок. Локальное распознавание СберБум Трек убирает зависимость музыкальных команд от облака; дайджест Mail сокращает объём переписки, который человеку приходится просматривать самому. Обе новости показывают прикладную ценность небольших, чётко ограниченных ИИ-функций, а не универсального обещания «ассистент сделает всё».

Мир: модели, продукты и капитал.

17 ноября —

OpenAI выпустила GPT-6.2 Astra и Sol.

OpenAI представила две модели: Astra для сложных профессиональных проектов и Sol для повседневной работы. Обе доступны через API и в соответствующих планах ChatGPT Work и Codex. Компания выделяет создание документов, таблиц, интерфейсов и выполнение многошаговых задач. Стандартные цены API за миллион входных/выходных токенов: Astra — $8/$40, Sol — $2/$10. Отдельно ChatGPT for Healthcare получил агентов для работы с медицинским ПО, улучшенный клинический поиск и более 20 плагинов. Партнёрство с медиаподразделением LVMH охватывает инструменты для редакций и представление их материалов в ChatGPT со ссылками.

Почему важно. Профессиональный результат требует не просто ответа, а документа нужного формата и работы с несколькими системами. GPT-6.2 и отраслевые подключения OpenAI сдвигают продукт в сторону завершённого рабочего процесса. Разделение Astra и Sol позволяет не оплачивать максимальную вычислительную сложность для каждого поручения; медицинские интеграции сокращают ручной перенос информации, но не превращают модель в самостоятельного врача.

18–19 ноября —

Google: Gemini 4 Neon для работы, Genie 4 для интерактивных миров.

Gemini 4 Neon — модель Google для программирования, анализа и научных задач с контекстом до миллиона токенов: в одном запросе можно обрабатывать большие наборы текстов, таблиц и мультимедийных материалов. Стандартные цены API — $3 за миллион входных и $15 за миллион выходных токенов. Другой продукт, Genie 4, вышел в Gemini API и AI Studio: по тексту или изображению он генерирует интерактивные миры 720p при 24 кадрах в секунду со звуком. Пользователь может перемещаться и менять окружение по ходу сессии.

Почему важно. Обычному ассистенту приходится дробить большой проект на фрагменты, а видеогенератору — заново создавать ролик после изменения замысла. Большой контекст Gemini 4 Neon помогает удерживать связанные материалы вместе. Genie 4 заменяет просмотр фиксированного видео управляемой средой: дизайнеры и разработчики могут быстрее проверять пространственные идеи и прототипы взаимодействия. Генерируемый мир при этом не равен физически достоверному симулятору.

16–20 ноября —

Gemini начинает действовать на Mac и расширяет бытовые поручения.

Google начала запуск управления компьютером в Gemini для Mac: ассистент работает с выбранными приложениями, файлами и сайтами. Первоначально функция доступна на английском в США подписчикам AI Plus, Pro и Ultra. Tasks в Gemini расширяет бесплатный доступ и многошаговые поручения; Connected Apps добавляют в США покупку продуктов у новых продавцов. Параллельно Google разделяет сохранение истории и использование чатов для обучения, вводит временные диалоги AI Mode и расширяет Gemini for Home ещё на шесть стран.

Почему важно. Раньше пользователь получал план действий, но сам переключался между приложениями и оформлял заказ. Управление компьютером и подключённые сервисы Gemini позволяют передать часть исполнения ассистенту. Разделение настроек истории и обучения снимает другой компромисс: хранить полезную переписку можно без согласия на использование разговоров для улучшения генеративных моделей. Региональные ограничения пока заметно сужают доступность новых сценариев.

---

20 ноября —

Anthropic закрыла IPO.

Разработчик Claude завершил первичное публичное размещение акций по $83. Валовые поступления IPO, включая продажу акций существующими держателями, составили около $93,5 млрд; самой Anthropic досталось около $91,6 млрд до комиссий и расходов. Отдельное одновременное частное размещение принесло около $26 млрд. Торги на Nasdaq под тикером ANT начались 19 ноября.

Почему важно. Обучение моделей и обслуживание пользователей требуют капитала задолго до того, как соответствующие мощности начнут окупаться. IPO Anthropic расширяет источники финансирования за пределы закрытых инвестиционных раундов. Одновременно публичный статус усиливает требования к финансовой отчётности: рынку будет проще сопоставлять рост использования Claude с затратами компании. Размер привлечения сам по себе не доказывает ни прибыльность бизнеса, ни будущую доходность акций.

16–17 ноября —

Cohere добавляет поиск по видео и бизнес-данным.

Cohere представила Video Intelligence в Compass: поиск и извлечение информации из видео и документов с фрагментами, расшифровками, описаниями изображения и временными отметками. Компания набирает партнёров для совместной разработки, а не объявляет общую доступность. Коннектор Crunchbase для North уже позволяет подключать сведения о компаниях и рынке; нужны действующая лицензия Crunchbase и единый вход SSO.

Почему важно. Корпоративные знания остаются не только в документах: ответ может находиться в середине записи встречи или в отдельной базе о компаниях. Поиск Cohere связывает вопрос с конкретным фрагментом

Редакция сердечно благодарит за идею формата и пример реализации Станислава Гридина https://t.me/gridstan

видео и подключёнными бизнес-данными, сокращая ручной просмотр и копирование. Для аналитиков и сотрудников поддержки результат становится полезнее, когда к ответу можно сразу открыть исходный момент записи или запись базы.

Безопасность и доверие.

16–20 ноября —

Anthropic сокращает сроки исправления уязвимостей; независимая оценка становится ближе к разработке.

Anthropic сократит срок согласованного раскрытия уязвимостей с 90 до 30 дней для критических и высокоопасных проблем и до 45 дней для остальных — для сообщений с 1 января 2027 года. Google и UCL представили HelpBench для оценки повседневных киберсоветов ИИ
. OpenAI заключила соглашения с Apollo Research, METR, Transluce и Goodfire
об оценках с глубоким доступом к внутренним моделям и системам. В карте GPT-6.2 компания прямо описывает рост кибервозможностей и ограничения контроля поведения.
Экспертные советники Еврокомиссии
призвали усиливать независимую оценку; заявление не является новым законом. В России
F6 связала атаки Rezet с панелью AnyDoor,
предположительно созданной с помощью ИИ. Salesforce пояснила возможность удаления записей через Claude
, если права пользователя и настройки Dispatch разрешают такие действия; Okta расширила управление доступом агентов к MCP-серверам
.

Почему важно. Между обнаружением уязвимости и её массовым использованием раньше могло оставаться больше времени; ИИ ускоряет повторное обнаружение и разработку атак. Короткие сроки Anthropic меняют темп исправлений. HelpBench и допуск внешних оценщиков OpenAI решают соседнюю проблему: убедительный ответ или хороший тестовый балл не показывают, насколько система опасна в реальной работе. Оценка конкретных советов и поведения во время разработки помогает обнаруживать риски до широкого применения. Примеры F6 и Salesforce показывают два практических следствия: автоматизация масштабирует не только полезные задачи, но и атаки, а доступ агента к данным превращается в возможность необратимого действия. Ограниченные по задаче права, которые развивает Okta, уменьшают масштаб возможного ущерба.

Инфраструктура и корпоративные агенты.

17–20 ноября —

Microsoft Ignite: агенты входят в корпоративный рабочий контур.

Microsoft расширила Copilot и Foundry: среди анонсов Ignite — общедоступные виртуальные машины Cobalt 200 и предварительные возможности подключения моделей и агентов Foundry к клиентским кластерам AKS. Replit добавила интеграции с Fabric и Teams и ранний доступ к публикации в Azure; Databricks Genie расширяет работу с корпоративными данными из инструментов Microsoft.
Hermes Enterprise на Microsoft Foundry
показывает агента с собственной рабочей идентичностью; направление ещё развивается. NVIDIA расширила доступ к Nemotron 3.5 через Foundry и сообщила о системах Vera Rubin NVL72 в Azure.

Почему важно. Самостоятельный чат-бот отделён от корпоративных данных, прав доступа и среды исполнения. Интеграции Ignite помещают создание приложений, анализ и действия агентов ближе к существующим системам компании. Для сотрудников уменьшается число переходов между программами; для ИТ-команд появляется общий контур размещения и администрирования. Собственная идентичность агента позволяет отделить его действия от действий человека, вместо передачи агенту всей учётной записи сотрудника.

16–18 ноября —

NVIDIA: научные агенты, процессоры Vera и локальные RTX Spark.

NVIDIA представила Nemotron for Science: модель Nemotron 3.5 Super VL, научные оценки, открытый химический датасет и рецепты дообучения для научных агентов. Стоечная платформа Vera CPU объединяет 192 процессора; системы партнёров ожидаются с начала 2027 года. Отдельно RTX Spark от ASUS, Dell и MSI поступили в продажу
с объединённой памятью до 128 ГБ для локального ИИ. Анонсированный Rubin HPC ожидается в 2028 году — не следует путать объявление с поставками.

Почему важно. Научному ИИ нужны специализированные данные и доступ к расчётам, а не только общие знания чат-бота. Nemotron for Science даёт исходные компоненты для такой адаптации; Vera ориентирована на требовательные вычисления и обмен данными. RTX Spark решает задачу другого масштаба: небольшая команда может запускать подходящие модели рядом со своими файлами, не отправляя каждый запрос облачному провайдеру. Объём памяти при этом не гарантирует приемлемой скорости любой модели.

16–18 ноября —

Dell расширяет AI Factory, AMD покупает Taalas.

Dell расширила серверные и стоечные варианты AI Factory на базе AMD и NVIDIA. AI Data Platform получает средства предоставления корпоративных данных агентам; OpenManage Rack Scale объединяет наблюдение за вычислениями, питанием и охлаждением.
Архитектура Dell с OpenAI Codex связывает программирующего агента с управляемыми репозиториями и данными — без обещания локальных весов OpenAI. 17 ноября AMD завершила покупку Taalas, разработчика специализированных чипов для исполнения моделей.

Почему важно. Покупка ускорителей не решает сама по себе проблемы доступа к данным, охлаждения и работы приложений. Комплекс Dell связывает эти части в варианты от отдельного узла до стойки, уменьшая объём самостоятельной сборки корпоративного ИИ. Покупка Taalas расширяет технологические варианты AMD для инференса: конкуренция может идти не только за универсальные GPU, но и за специализированное исполнение моделей. Конкретные будущие продукты и экономия пока не определены сделкой автоматически.

18–21 ноября —

AWS связывает документы, речь и совместную работу агентов.

Общедоступный Textract Insights извлекает данные из документов по описанию нужных полей, показывает уверенность и привязку к страницам. AWS Quick получил создание процессов в чате и совместную работу специализированных агентов; Amazon Transcribe — синхронный API, возвращающий расшифровку сразу в ответе. DevOps Agent расширил работу через MCP-клиенты, Slack и Teams. SageMaker JumpStart предлагает сравнение более 70 моделей по качеству, скорости и стоимости; Partner Central добавил автоматизацию задач партнёрских продаж.

Почему важно. При автоматизации бизнеса время тратится не только на рассуждение модели, но и на подготовку входных данных, поиск ошибок и соединение этапов процесса. Обновления AWS сокращают эти промежутки: документы возвращают проверяемые по странице значения, аудио — готовую расшифровку, а агенты распределяют работу.

Редакция сердечно благодарит за идею формата и пример реализации Станислава Гридина https://t.me/gridstan

Сравнение моделей помогает связать выбор поставщика с конкретной ценой и задержкой, а не только с рекламным рейтингом.

Открытые модели и инструменты.

18–19 ноября —

Cloudflare открыла Clairic, OVH — несколько вариантов Olala.

Cloudflare опубликовала открытые веса обновлённой Clairic и облегчённой Clairic-flash на 9 млрд параметров для извлечения данных из документов, изображений и PDF; модели доступны также в Workers AI. OVHai LLM выпустила Olala для Apple Silicon в формате MLX, SFT-чекпойнт до финального выравнивания и специализированную модель суммаризации. MLX-варианту нужен указанный разработчиком форк mlx-engine; исследовательский SFT-чекпойнт не является готовым безопасным ассистентом.

Почему важно. Облачный API удобен, но привязывает обработку документов к чужой инфраструктуре и тарифам. Доступные веса Clairic и Olala позволяют выбирать локальное размещение и адаптировать модель под конкретный тип данных. Специализация на извлечении или суммаризации может оказаться полезнее универсальной беседы. Различия между форматами, лицензиями и стадиями обучения существенны: открытые веса не означают автоматически одинаковую свободу использования или одинаковое качество поведения.

16–20 ноября —

Инструменты агентов экономят память, токены и работу инженеров.

Google открыла предварительный Tool Retrieval в Gemini Interactions API: агент загружает описания нужных инструментов по запросу, вместо передачи всей библиотеки каждый раз. Superfluid Fleet объединяет локальные компьютеры в одну точку обслуживания моделей; Together Storage Accelerator кеширует данные на GPU-серверах. Devin for Databricks Migrations автоматизирует преобразование и проверку кода при миграции; Unity Gateway Insights ищет перерасход токенов и сбои в трассировках агентов. В открытом стеке вышел vLLM 0.34.0; Hugging Face представила адаптеры под агентские среды.

Почему важно. Дорогой агент может тратить ресурсы на лишние описания инструментов, повторное чтение данных и одинаковые ошибочные циклы. Tool Retrieval и кеширование уменьшают такую накладную работу; наблюдаемость Databricks помогает обнаружить её источник. Распределение локальных сессий и адаптация к среде исполнения повышают отдачу имеющегося оборудования. Поэтому стоимость выполненной задачи определяется не одной моделью: устройство всего рабочего контура способно заметно изменить результат.

AI Coding Newsletter · Telegram · LLM Notes

Что отслеживать

- **16–20 ноября — Яндекс подключает Алису и внешних агентов к сервисам.** Расширение предварительного доступа и доступных действий в Яндекс 360, состав и условия каталога плагинов Алисы.
- **20 ноября — MWS AI открыла CoSpaDi для сжатия языковых моделей.** Код, наборы испытаний и дальнейшие сравнения в объявлении MWS AI.
- **17 ноября — Российский бытовой ИИ: локальный голос и разбор почты.** Поддерживаемые офлайн-команды СберБум Трек и доступность функций Mail на других платформах.
- **17 ноября — OpenAI выпустила GPT-6.2 Astra и Sol.** Доступность и цены GPT-6.2, реальные ограничения из карты безопасности и расширение медицинских интеграций.
- **18–19 ноября — Google: Gemini 4 Neon для работы, Genie 4 для интерактивных миров.** Практическое качество длинных задач Gemini 4 Neon и ограничения, стоимость и длительность сессий Genie 4 API.
- **16–20 ноября — Gemini начинает действовать на Mac и расширяет бытовые поручения.** Расширение географии Gemini для Mac и набор реальных действий Tasks.
- **20 ноября — Anthropic закрыла IPO.** Последующие финансовые раскрытия и использование средств в материалах Anthropic об IPO.
- **16–17 ноября — Cohere добавляет поиск по видео и бизнес-данным.** Переход Video Intelligence от партнёрской разработки к более широкой доступности и набор данных коннектора Crunchbase.
- **16–20 ноября — Anthropic сокращает сроки исправления уязвимостей; независимая оценка становится ближе к разработке.** Вступление в силу сроков Anthropic, результаты встроенных внешних оценок OpenAI и обновления HelpBench.
- **17–20 ноября — Microsoft Ignite: агенты входят в корпоративный рабочий контур.** Переход предварительных функций Foundry и AKS к общей доступности и развитие Hermes Enterprise.
- **16–18 ноября — NVIDIA: научные агенты, процессоры Vera и локальные RTX Spark.** Научные результаты и материалы Nemotron for Science, начало поставок Vera CPU и реальные конфигурации RTX Spark.
- **16–18 ноября — Dell расширяет AI Factory, AMD покупает Taalas.** Сроки доступности компонентов Dell AI Factory и применение технологии Taalas в продуктах AMD.
- **18–21 ноября — AWS связывает документы, речь и совместную работу агентов.** Качество извлечения Textract Insights на сложных документах и доступность функций совместной работы Quick.
- **18–19 ноября — Cloudflare открыла Clairic, OVH — несколько вариантов Olala.** Лицензии и поддерживаемые форматы Clairic, развитие Olala MLX и точность Olala summarise на длинных документах.
- **16–20 ноября — Инструменты агентов экономят память, токены и работу инженеров.** Ограничения Tool Retrieval, распределение нагрузки Superfluid Fleet и результаты реальных миграций с Devin.
- Российские рабочие сервисы: какие действия выйдут из предварительного доступа в MCP-сервере Яндекс 360.
- Экономику моделей: итоговую стоимость завершённого поручения, а не только токена, у GPT-6.2 и Gemini 4 Neon.
- Реальное исполнение: расширение возможностей и географии Gemini для Mac и развитие Hermes Enterprise на Foundry.
- Внешнюю оценку: какие ограничения и результаты будут опубликованы в рамках соглашений OpenAI с независимыми исследователями.
- Локальный ИИ: доступность аппаратных конфигураций RTX Spark и сохранение качества после сжатия CoSpaDi.

Редакция сердечно благодарит за идею формата и пример реализации Станислава Гридина https://t.me/gridstan